8887-EN-01
DTIC

APPLICATION OF GEOSTATISTICAL METHODS AND WAVELETS TO THE ANALYSIS OF HYPERSPECTRAL IMAGERY AND THE TESTING OF A MOVING VARIOGRAM

Third Interim Report (RSSUSA - 5/3)

Dr Margaret A. Oliver

November 2000 to January 2001

United States Army

ENVIRONMENTAL RESEARCH OFFICE OF THE U.S. ARMY

London, England

CONTRACT NUMBER - N68171-00-M-5508

Contractor - Approved for Public Release; distribution unlimited

20010502 048

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503.

1. AGENCY USE ONLY (Leave blank)

2. REPORT DATE
16.02.01

3. REPORT TYPE AND DATES COVERED
Interim Nov 2000 – Jan 2001

4. TITLE AND SUBTITLE
Application of geostatistical methods and
wavelets to the analysis of hyperspectral
imagery and the testing of a moving variogram

5. FUNDING NUMBERS
N68171
00-M-5508

6. AUTHOR(S)
Dr Margaret A Oliver

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
University of Reading, Whiteknights,
Reading, RG6 2AH, UK

8. PERFORMING ORGANIZATION REPORT NUMBER
(RSSUSA-5/3)

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
USARDSG-UK, Environmental Sciences Branch
Edison House, 233 Old Marylebone Road,
London, NW1 5TH, UK

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES
Interim Report:
Summary of work to date

12a. DISTRIBUTION/AVAILABILITY STATEMENT
No limitation on distribution/availability

12b. DISTRIBUTION CODE

13. ABSTRACT (Maximum 200 words)

This is the third report of the project. It is a short one, but it is a stand-alone piece of work. It is an aide memoire for sampling. It embraces both design-based sampling, which is based on classical statistics, and model-based sampling which is underpinned by geostatistics. This work is a guide for sampling in the field or pixels from images. It starts with what the user should consider before sampling, i.e. the target population, the sample support (volume of sample), the individuals, what the data are to be used for, what kind of predictions are required. Based on the kind of predictions required the user will decide either to have a sample design for design-based estimation or one for model-based prediction.

14. SUBJECT TERMS
Keywords: sampling, design-based, model-based, variogram, kriging variances

15. NUMBER OF PAGES

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT

18. SECURITY CLASSIFICATION OF THIS PAGE

19. SECURITY CLASSIFICATION OF ABSTRACT

20. LIMITATION OF ABSTRACT

Attachment 1 Page 1of2

Aide mémoire – Spatial sampling

Introduction

The land surface, the materials of which it is composed and the environment more generally are continuous. In general measurements or observations can be made on only small portions of them, i.e. on samples, because of the large areas involved. For example, in a single agricultural field there is an infinite number of potential sampling points. Samples intended to represent the areas from which they are drawn must be planned with care. The information from a sample location should represent a surrounding area, the extent of which we might not know. Since many environmental properties vary locally in a complex and erratic way the values from a single sampling point include a sampling effect. To increase the information from a sampling location so that it is representative a bulked sample can be taken, and provided that the property is additive the measurement made on it will equal the regional mean apart from sampling error.

At the outset consider the use that will be made of the sample information. For instance, will the mean values of the properties observed for the entire area or for strata within the area be used to predict at unsampled places? Or will the information be used to predict locally, either using mathematical interpolators or geostatistical ones. For either of the latter the sample data must be spatially autocorrelated for them to have any merit.

This aide lists the matters that must be considered and resolved in planning sampling of a geographic region, which for present purposes we treat as two-dimensional.

Defining the target

The domain

The domain is the region of interest. Circumscribe it by a boundary on a map so that every point can be assigned to the domain or not with certainty. The domain may comprise a single parcel of land or several. Denote it by *D*.

Support

The support is the area or volume of material on which you make measurements. It has size and shape, and may have orientation. In remote sensing it is the `footprint' of the pixel; in vegetation surveys it is the quadrat; in soil survey it is the core of soil taken from the ground. Cores of soil may be taken from areas larger than the cross-section of the cores and bulked for analysis in the laboratory. In these cases the supports are the larger areas.

In any one survey define the support and keep it constant throughout.

The population and units

Within D are units that have the dimensions of the supports. In a remote image their number is finite though large. In soil survey they are so many that they may be regarded as infinite. Define them by their spatial coordinates and their spatial extents. Together they comprise the population. The terms 'population' and 'units' may be used to refer to the values of a variable of the supports.

The target

Within D there may be only certain kinds of terrain or land use that are of interest, e.g. only dry land (not water), only farm land (not towns, not parks, not golf-courses, etc.). The units falling in these classes constitute the target population. The others do not belong.

Samples

Whole populations cannot be measured in ground survey; you can measure only subsets of the units that comprise them. Such a subset of units is a sample.

Typically you will want two characteristics in a sample – accuracy and reliability. The first means that a sample represents the population without bias, i.e. any value that we obtain from a sample will be as likely to exceed the true value of the population as it will be to fall short. The second implies that repeated sampling will give sensibly the same result. It is measured by the estimation variance or standard error of the mean, s.e.

These characteristics can be assured by the sampling design in which there is sufficient randomness.

Notation

We adopt the following basic notation.

D denotes the domain.

$|D|$ is the area of D.

z is the variable of interest.

Z is a random variable.

$Z(\mathbf{x})$ is a random process, random field, stochastic process, in which Z may take any one of two or more values at random at each point $\mathbf{x}$ in D.

N is the size of the sample in D, i.e. the number of units in it.

D_k denotes the kth subdivision of D, of which there may be K.

n_k is the number of units in a sample of D_k.

μ denotes the mean of z in D.

$\bar{z}$ is the mean of the N data drawn from D.
σ^2 is the variance of z in D.

$s^2 = \hat{\sigma}^2$ is the estimate of σ^2 from the N data.

$s^2(D)$ is the estimation variance of μ in D.

$s(D)$ is the standard error of μ.

$\mathbf{h}$ denotes the lag separating two places, and is a vector in two dimensions; $|\mathbf{h}|$ is the distance component of the lag.

$\gamma(\mathbf{h})$ signifies the semivariance at lag $\mathbf{h}$.

λ_i are the kriging weights.

Sampling designs for design-based estimation

This is essentially the classical statistical approach to sample design and prediction.

Simple random sampling

In simple random sampling N units are chosen with equal probability from the target population. The result is unbiased, and the estimation variance $s^2(D)$ is given by s^2/N.

If there is any spatial correlation at the working scale then this is inefficient in the sense that the same estimation variance could be achieved with a smaller sample by a better design.

Stratified random sampling

Divide the region into strata, D_k, k=1,2, . . ., K, and represent each by a few units, ideally two, chosen at random independently. The sizes n_k may be chosen in proportion to the areas of the D_k, $|D_k|$, if they are not equal.

If other sizes are chosen then the mean in D may be calculated as the weighted average of the individual stratum means with weights proportional to the $|D_k|$. The estimation variance of stratified sampling depends on the variance within the strata, or the pooled within stratum variance. In the presence of spatial dependence the latter is less than the total variance in the population, and so stratified sampling is more efficient than simple random sampling.

The estimation variance is given by

$$s^2(D)_{\text{stratified}} = \sum_{k=1}^{K} w_k^2 s^2(D_k),$$

where $s^2(D_k)$ is the estimation variance within stratum D_k, and w_k is the weight assigned to the stratum. The weights should sum to 1 to avoid bias.

There are numerous ways in which this general scheme can be elaborated according to what you know of the region and the variation within it. For example the strata could have unequal spatial extents as in classification. In this case the different areas are taken into account through the weight w_k, such that

$$w_k = \frac{\text{area of stratum } k}{\text{total area}} \quad .$$

Systematic sampling

Sampling is usually most efficient when done on a regular grid. It has two disadvantages:

(1) it provides no ready estimate of the variance;

(2) it may lead to biased estimates of the mean.

The first arises because once the origin and orientation of the grid are decided there is no further randomization possible. It is not easily overcome, but the estimation variance may be approximated by methods such as Yates's balanced differences.

The second, bias, can happen where there is trend or periodicity in z in the region. Periodicity is usually evident, and if it is then you can choose an interval and orientation that will be out of tune with it. Alternatively, choose a non-aligned scheme in which each sampling point on the grid is offset from its node by a random distance along its row and down its column according to a rule.

Sample size

The size of sample N may depend on the budget or the tolerance, i.e. error that can be tolerated in the estimate from the survey. If the budget is fixed then choose a stratified scheme to minimize the error for that budget.

If the error is specified as $s(D)$ then for simple random sampling

$$N = s^2 / s^2(D) \,,$$

The formula for stratified sampling is more elaborate.

You usually do not know s^2 in advance, and so choosing N is problematic. Therefore sample in stages, starting with a sparse design that can be intensified as necessary. At each stage calculate the estimation variance to see whether it meets the tolerance. If it does then stop; otherwise intensify the sampling and recompute the estimation variance as the next stage.

Geostatistical (model-based) sampling design and prediction

Geostatistics is used to estimate local values rather than regional ones, i.e. to predict. It is based on the assumption that z in the real world is a realization of the random process $Z(\mathbf{x})$. For this reason there is no need to randomize the sampling, and grid sampling is preferred because of its efficiency.

Geostatistical prediction (kriging) requires a model of the correlation structure, expressed either as a covariance function, or rather more generally as a variogram. Like the variance in design-based estimation, these functions are not known *a priori* and must be estimated from sample data. Sampling must therefore serve two purposes:

(1) estimation and modelling of the variogram, and

(2) local prediction once the variogram has been estimated and modelled.

To satisfy item (1) sampling must be sufficient to estimate the semivariances precisely. It must also be dense enough to estimate the spatial characteristics of the variation, such as correlation range and general form of the variogram.

Sampling for item (2) will depend either on the budget or on the tolerable error of local predictions and the variogram.

Sampling to estimate the variogram

Nested sampling and analysis

Start with nested sampling and a hierarchical analysis of variance of the sample data if you know nothing of variation in the region. Choose five or six sampling intervals in geometric progression from the smallest lag distance of interest to the largest. Choose the angular separations at random. Replicate at the longer distances to give sufficient degrees of freedom in the analysis of variance to estimate the components. Expect to have a total sample, N, of about 100. Figure 1 shows the kind of sampling plan to aim for.

Accumulate the components of variance to estimate $\gamma(|\mathbf{h}|)$ at the distances of the design and draw a crude variogram with the logarithm of $|\mathbf{h}|$ on the abscissa as in Figure 2.

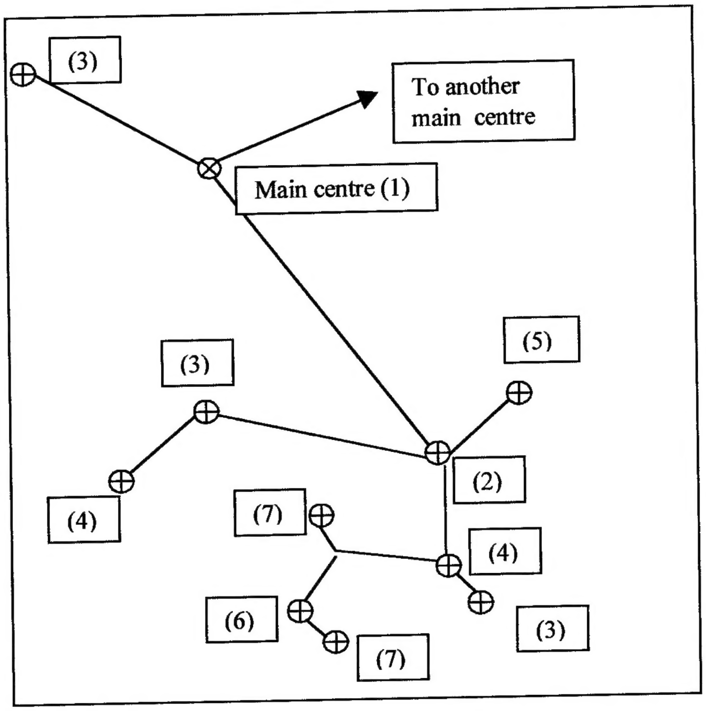

Figure 1. The plan of sampling for one main centre in a nested survey with 7 stages. The stages in the hierarchy are given for each sampling site.

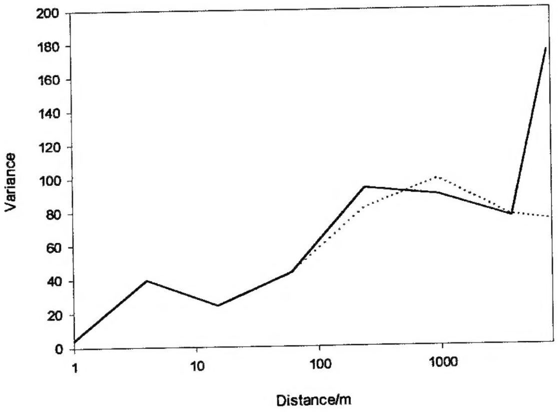

Figure 2. The accumulated components of variance from a hierarchical analysis of variance giving a first approximation to the variogram.

Such a result this can be used to identify the range of distance within which most variance occurs and to plan further sampling to estimate the conventional variogram.

If all the variance appears to occur within the smallest distance of interest then local prediction is not feasible. So stop! Figure 3 shows an example of a pure nugget reconnaissance variogram. All of the variation is occurring within the shortest sampling interval.

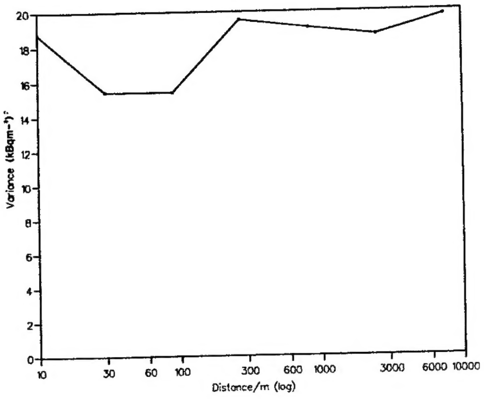

Figure 3. A pure nugget reconnaissance variogram from a nested survey.

Estimating the variogram parameters

Use the result from the hierarchical analysis above or other knowledge of the variation in *D* to estimate semivariances, $\gamma(\mathbf{h})$, at several lags, **h**, within the correlation range. Design a scheme with approximately 100 to 150 sampling points if the variation appears isotropic. If a square grid with this number gives you sufficient estimates of $\gamma(\mathbf{h})$ within the correlation range then use it. If not then cluster the sampling in some way. Intensify sampling around a subset of grid nodes, bearing in mind that you are likely to want a grid for kriging later. Alternatively, sample in clusters with a range of sampling distances between locations, and spread the clusters evenly over *D* so that the

Do not cluster sampling in parts of *D* that you know or suspect to have unusually large values of *z* (as you might in mineral surveys or pollution studies) or unusually small ones (as in studies of deficiency diseases). This will result in bias.

Compute the sample variogram and plot the result. If the estimated values fall close to a smooth curve then choose an authorized model to describe it, estimate its parameters, and proceed to kriging.

If there is too much scatter to identify a plausible function then increase the sampling, either by intensifying the grid or by adding clusters, and recompute the variogram. Repeat until a smooth form is identifiable.

If the variation is anisotropic and you wish to model the anisotropy then you must expect to sample at 200 points or more.

Kriging

In kriging Z at an unknown point $\mathbf{x}_0$ minimize the prediction variance

$$\sigma^2(\mathbf{x}_0) = 2\sum_{i=1}^{n} \lambda_i \gamma(\mathbf{x}_0 - \mathbf{x}_i) - \sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i \lambda_j \gamma(\mathbf{x}_i - \mathbf{x}_j) \,, \qquad (1)$$

where $n << N$ is the number of sampling points near to the target point $\mathbf{x}_0$. The quantities $\gamma(\mathbf{x}_i - \mathbf{x}_j)$ and $\gamma(\mathbf{x}_0 - \mathbf{x}_j)$ depend on the separations $\mathbf{x}_i - \mathbf{x}_j$ and $\mathbf{x}_0 - \mathbf{x}_j$; the larger these are the larger is $\sigma^2(\mathbf{x}_0)$.

The maximum value of $\sigma^2(\mathbf{x}_0)$ is minimized by sampling on a regular grid. A triangular grid is usually the most efficient, but rectangular grids are almost as good (Figure 3a), and as they are easier to lay out and document they are preferred. If variation is isotropic then use a square grid.

If the budget is fixed then sample as intensely as it permits. If a maximum tolerance is specified, say s_{Kmax}, then solve the kriging system for a range of sampling intensities (grid intervals) and plot the kriging variance (or its square root, the kriging error) on the ordinate against the grid interval on the abscissa. Connect the points by a smooth line, Figure 3. From s^2_{Kmax}, or s_{Kmax}, draw a horizontal line until it meets the curve, and from that intersection drop a perpendicular. The value at which the perpendicular cuts the abscissa is the required sampling interval, Figure 3b.

Determine the number of cores in bulked samples similarly. Compute the estimation variances using Equation (1) for equispaced sampling configurations and sample sizes from 4 to about 50 and join the values to form a curve (Figure 4). Draw a horizontal line at the maximum tolerable variance, and drop a perpendicular from the point at where it intercepts the curve to the abscissa. The value on the abscissa is the optimum size of sample.

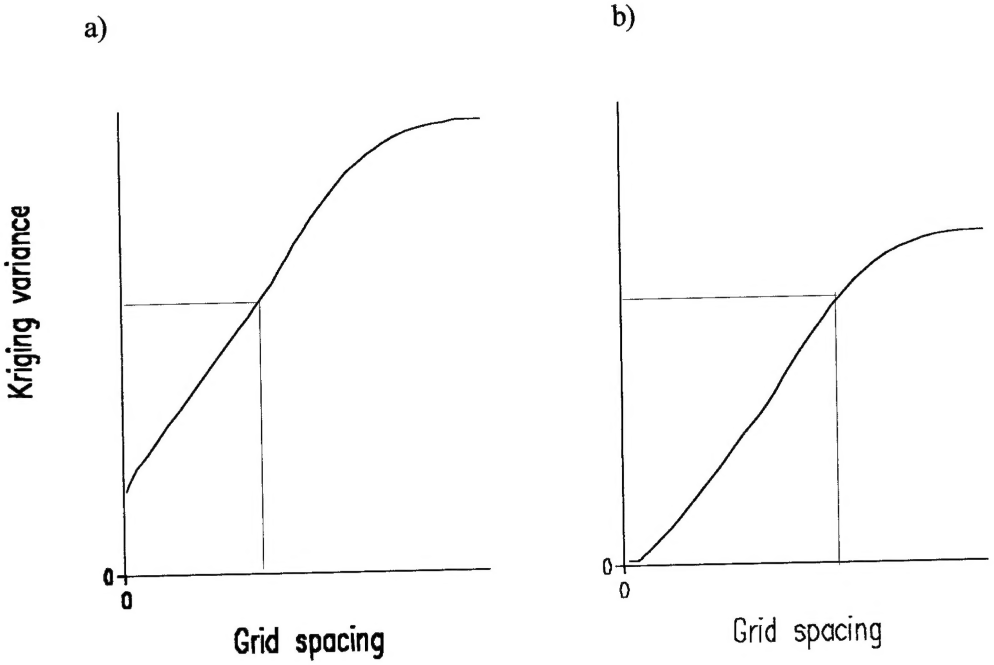

Figure 3. Kriging variances from (a) punctual kriging, and (b) block kriging.

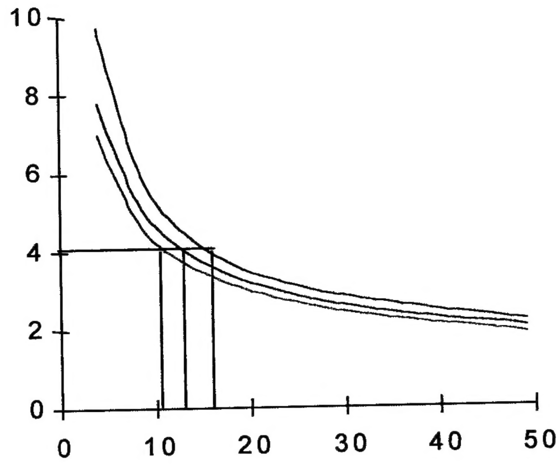

Figure 4. Graphs of standard error plotted against sample size for bulking from 4, 9, 16, 25, 36 and 49 cores, and for three different sample supports.